

Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the “Confirmation of Authorship” file.

As corresponding author I, Angeliki Papadimitriou, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature _____ Date _____

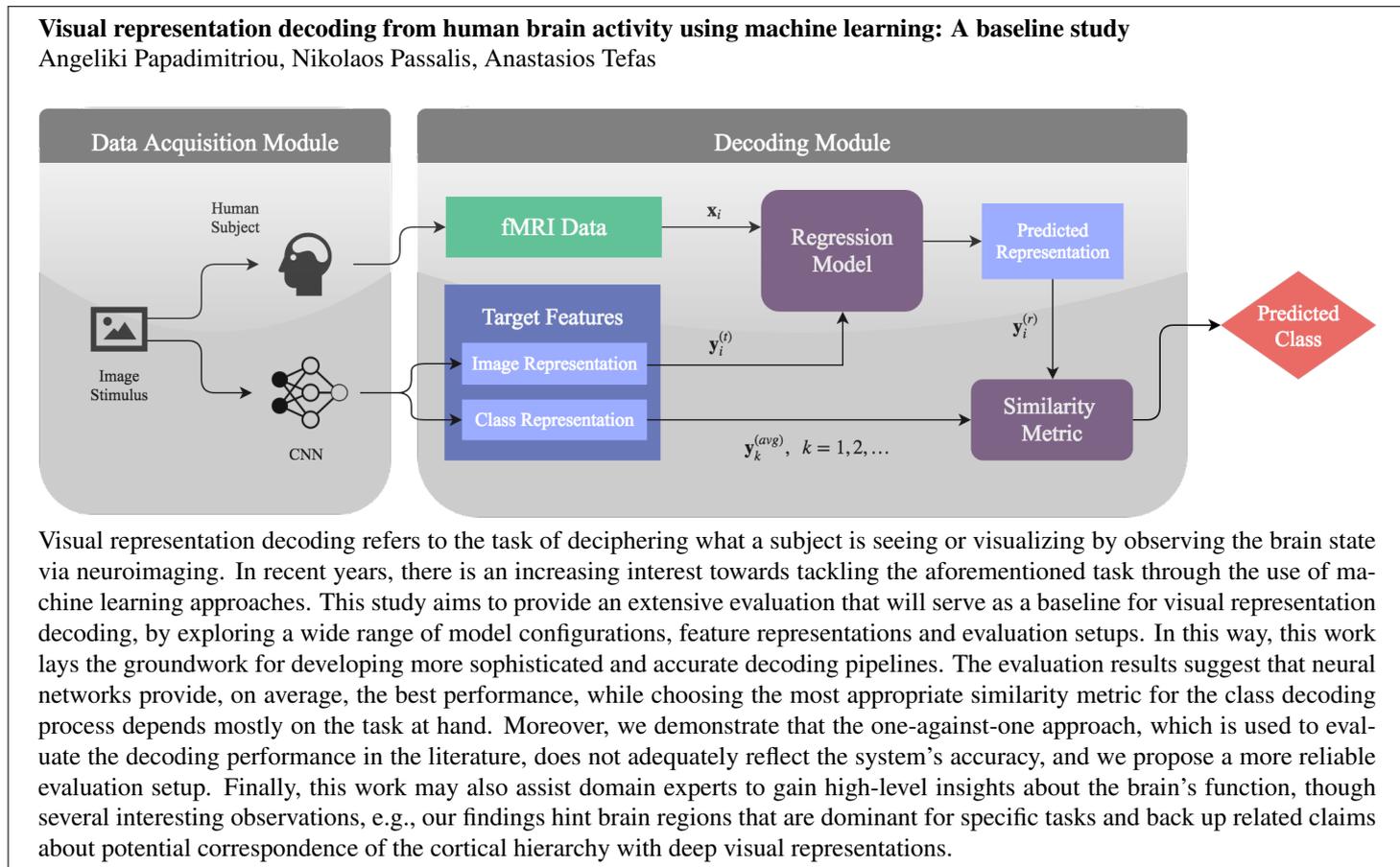
List any pre-prints:

Relevant Conference publication(s) (submitted, accepted, or published):

Justification for re-publication:

Graphical Abstract (Optional)

To create your abstract, please type over the instructions in the template box below. Fonts or abstract dimensions should not be changed or altered.



Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- Baseline evaluation for visual representation decoding from brain activity.
- A new, more reliable evaluation setup is proposed.
- Exhaustive evaluation of various machine learning methods for the decoding.
-
-



Visual representation decoding from human brain activity using machine learning: A baseline study

Angeliki Papadimitriou^{a,**}, Nikolaos Passalis^a, Anastasios Tefas^a

^aArtificial Intelligence and Information Analysis Laboratory, Aristotle University of Thessaloniki, Thessaloniki 541 24, Greece

ABSTRACT

Visual representation decoding refers to the task of deciphering what a subject is seeing or visualizing by observing the brain state via neuroimaging. In recent years, there is an increasing interest towards tackling the aforementioned task through the use of machine learning approaches. This study aims to provide an extensive evaluation that will serve as a baseline for visual representation decoding, by exploring a wide range of model configurations, feature representations and evaluation setups. In this way, this work lays the groundwork for developing more sophisticated and accurate decoding pipelines. The evaluation results suggest that neural networks provide, on average, the best performance, while choosing the most appropriate similarity metric for the class decoding process depends mostly on the task at hand. Moreover, we demonstrate that the one-against-one approach, which is used to evaluate the decoding performance in the literature, does not adequately reflect the system's accuracy, and we propose a more reliable evaluation setup. Finally, this work may also assist domain experts to gain high-level insights about the brain's function, though several interesting observations, e.g., our findings hint brain regions that are dominant for specific tasks and back up related claims about potential correspondence of the cortical hierarchy with deep visual representations.

© 2021 Elsevier Ltd. All rights reserved.

1. Introduction

Neural decoding is the process of cracking the code of the brain encoding mechanism by observing the brain activity of subjects executing a specific task. Brain activity can be recorded via functional neuroimaging methods that are capable of capturing the signal changes over time. Such methods are Electroencephalography (EEG) (Niedermeyer and da Silva, 2005; Arunkumar et al., 2017; Pacola et al., 2017; Patel et al., 2018) and functional Magnetic Resonance Imaging (fMRI) (Huettel et al., 2004; Plumpton, 2014; Amin et al., 2017) can be used for various tasks and have provided data for several neural decoding endeavors in the literature, regarding tasks ranging from walking (Gwin et al., 2010; Presacco et al., 2011) and dancing (Cruz-Garza et al., 2014) to sleep (Tagliazucchi and Laufs, 2014; Horikawa et al., 2013).

Visual representation decoding is a special case of neural decoding that refers to the task of predicting the *visual stimulus*

that elicited the observed brain activity. This stimulus is encoded as an internal neural representation by the subject. The goal is to recover this representation by partially observing the brain state using a neuroimaging method, e.g., EEG or fMRI, allowing for classifying or even reconstructing the visual stimulus that elicited the observed activity. The term *generic decoding* refers to the additional property that allows for decoding representations of images that belong to categories for which the model has not been trained. Hence, it is desirable to design a model that not only learns the mappings of fMRI data to images, but can also competently generalize to recognize an arbitrary stimulus.

To this day, the way the brain encodes and organizes visual information remains a partially understood -yet immensely complex- process and an area of active research in neuroscience (Wen et al., 2017; Horikawa and Kamitani, 2017; Güçlütürk et al., 2017). The internal representations of the human brain are not known and, therefore, it is impossible to directly train a decoding model that translates brain activity signals to visual content. To overcome this limitation, an intermediate step is necessary to project both the visual content (im-

^{**}Corresponding author:
e-mail: akpapadim@csd.auth.gr (Angeliki Papadimitriou)

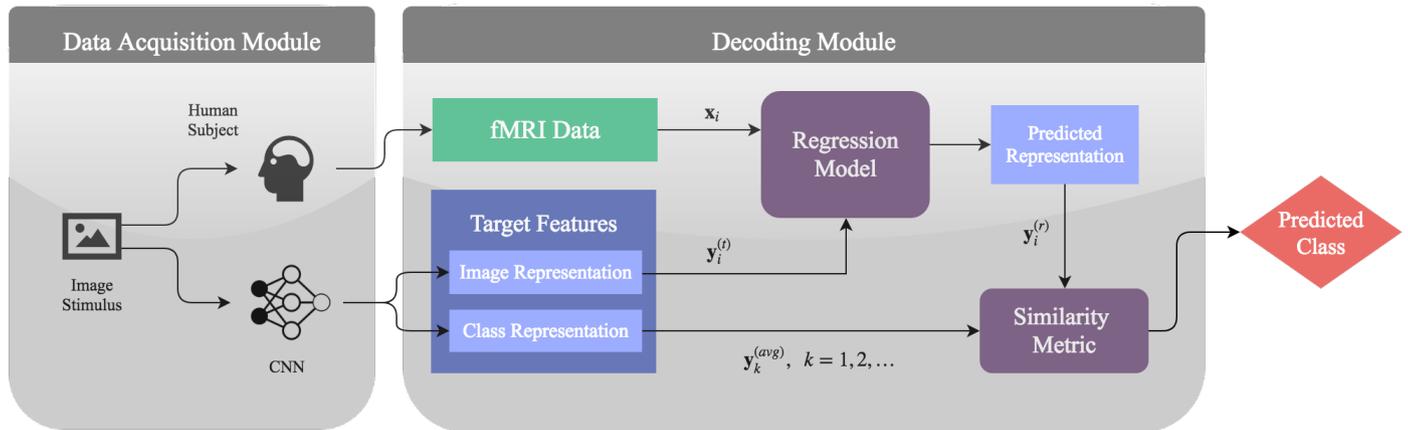


Fig. 1. Generic visual representation decoding from human brain activity data

ages) and the brain activity data to the same feature space and successfully decode the neural activity. The work by Horikawa and Kamitani (2017) hinted that the features extracted from various levels of a deep Convolutional Neural Network (CNN) are tightly correlated with the brain activity observed from various brain regions. Hence, CNN features can serve as the required (and appropriately regularized) intermediate neural representation that will assist a machine learning model (Nasrabadi, 2007) to recognize patterns in the observed brain signals and ultimately predict what a subject actually sees or imagines (Wen et al., 2017).

In recent years, there is an increasing interest in visual representation decoding with several methodologies being proposed (Horikawa and Kamitani, 2017; Wen et al., 2017; Güçlütürk et al., 2017; Wen et al., 2018). However, to the best of our knowledge, there exists no comprehensive assessment on the effect of different combinations of models, metrics and hyper-parameters on the decoding performance. The main contribution of this paper is to provide a solid baseline for the task of decoding generic visual representations from brain activity. Therefore, we extensively evaluate different design choices for several parts of the decoding pipeline proposed by Horikawa and Kamitani (2017), such as target features, regression models and similarity metrics. Moreover, we argue that the one-against-one approach, which is used to evaluate the model’s performance in the literature, does not adequately reflect the system’s accuracy, since it solves a significantly easier binary problem than the general (one-against-all) decoding problem. This is also reflected by the high chance level of 50% of the evaluation setup proposed by Horikawa and Kamitani (2017). Instead, we propose and evaluate the proposed approaches using a more realistic one-against-all setup, which more reliably reflects the accuracy of a neural decoding system. This paper also provides a reference point to develop more sophisticated and accurate decoding pipelines. The code used to perform the experiments is available on-line at `CODE-AVAILABLE-UPON-ACCEPTANCE`, allowing for easily reproducing the performed experiments.

The rest of this paper is structured as follows. The methodology applied in the present study, along with the machine learning models and evaluation setups are described in Section 2,

while the used dataset, experimental protocols and extensive evaluation results are presented in Section 3. Finally, conclusions are drawn in Section 4.

2. Generic Brain Activity Decoding

We follow the generic brain activity decoding pipeline proposed in (Horikawa and Kamitani, 2017). The employed pipeline is summarized in Fig. 1. First, a subject performs a visual task, while his/her brain activity is monitored using fMRI. Also, a CNN is used to extract a feature representation from the corresponding images (as they are presented to the subject). These representations can serve as a proxy for decoding the human brain activity for specific visual tasks, i.e., recognizing the class of an image shown to or visualized by a subject. To decode the brain activity, a regression model is used to directly predict the representation of the stimulus image using as sole input the measured fMRI signals. Subsequently, the class of the stimulus is deduced by comparing the decoded representation to a set of prototype class representations vectors. Note that this regression-based pipeline can be effectively used to infer the category/class for images which were *never presented* to the subject and/or regression model during the training.

The decoding pipeline is formally defined as follows. Let $\mathbf{x}_i \in \mathbb{R}^N$ be a N -dimensional feature vector that is extracted from the measured fMRI signals when the i -th experiment is performed (Horikawa and Kamitani, 2017). Also, the notation $\mathbf{y}_i^{(t)} \in \mathbb{R}^L$ is used to denote the representation extracted from a layer of a CNN, when the image used for the i -th experiment is fed to the network, while L is the dimensionality of this representation. Apart from the image representations, that are directly extracted from the CNN, prototype class representations are compiled for each class. This allows for generic class decoding for classes that were never presented to the model/subject during the training. The prototype class representation for the k -th class is defined as:

$$\mathbf{y}_k^{(avg)} = \frac{1}{|\mathcal{R}_k|} \sum_{\mathbf{y} \in \mathcal{R}_k} \mathbf{y} \in \mathbb{R}^L, \quad (1)$$

where \mathcal{R}_k is the set of CNN representations extracted from the images that belong to class k and $|\mathcal{R}_k|$ is the size of this set.

A machine learning model $f_{\mathbf{W}}(\mathbf{x})$ is then used to regress the image representation $\mathbf{y}_i^{(t)}$ using as input the corresponding brain activity \mathbf{x}_i . The notation \mathbf{W} is used to indicate the parameters of the employed model. The output of this model is calculated as $\mathbf{y}_i^{(r)} = f_{\mathbf{W}}(\mathbf{x}_i) \in \mathbb{R}^L$ and can be used to infer the class of the corresponding image by measuring the similarity of $\mathbf{y}_i^{(r)}$ with each of the class representations $\mathbf{y}_k^{(avg)}$. Therefore, the predicted class k^* of the object the subject sees or imagines is calculated as:

$$k^* = \arg \max_k S(\mathbf{y}_i^{(r)}, \mathbf{y}_k^{(avg)}), \quad (2)$$

where $S(\mathbf{y}_1, \mathbf{y}_2)$ is an appropriately defined similarity metric between two vectors $\mathbf{y}_1 \in \mathbb{R}^L$ and $\mathbf{y}_2 \in \mathbb{R}^L$. Note that (2) essentially describes a Nearest Centroid Classifier (NCC) (Schütze et al., 2008). The CNN representations $\mathbf{y}_i^{(t)}$ are only needed during the training process, as also shown in Fig. 1. Note that the image representations are not needed during the inference (test), since the class of the object that a subject sees or imagines can be inferred using only the (precomputed) class representation vectors.

2.1. Regression models

As mentioned in Section 1, the objective of this baseline study is, inter alia, to assess the performance of various machine learning models for the task of decoding the brain representations to the target feature vectors. In the present study, four different regression models are evaluated. The employed models are introduced below:

1. **k-Nearest Neighbor Regression (kNN):** The k nearest neighbors of a test sample are used in k-Nearest Neighbor Regression to infer the output representation by averaging the (known) target vectors of its neighbors in the train set (Hastie and Tibshirani, 1996). To take into account the proximity of each training sample to the current sample, the contribution of each neighbor is weighted according to its similarity to the current input.
2. **Linear Regression (LR):** Linear Regression a well-known and widely used regression model (Domingues et al., 2010; Cui et al., 2018; Yuan et al., 2012). The output of LR is calculated as: $f_{\mathbf{W}}(\mathbf{x}) = \mathbf{W}_{lr}\mathbf{x} + \mathbf{b}_{lr}$ (Mosteller and Tukey, 1977), where $\mathbf{W}_{lr} \in \mathbb{R}^{L \times N}$ is the matrix that contains the model’s parameters and $\mathbf{b}_{lr} \in \mathbb{R}^L$ is a vector that contains the independent terms. The mean squared error between the target representation and the output of the model is used for training:

$$\mathcal{L}_{mse} = \frac{1}{2N} \sum_i \|f_{\mathbf{W}}(\mathbf{x}_i) - \mathbf{y}_i^{(t)}\|_2^2, \quad (3)$$

where the notation $\|\mathbf{x}\|_2$ is used to denote the l^2 norm of a vector \mathbf{x} and N is the number of samples used for training the regression model. The model can be also *regularized* to avoid overfitting phenomena. For example, in Ridge Regression (RR) (Hoerl and Kennard, 1970), the l^2 norm of the regression parameters is used to this end.

3. **Kernel Regression (KR):** Kernel Regression is a powerful non-linear variant of LR. In KR the data are first projected into a higher-dimensional space, where they can be better separated (Shawe-Taylor et al., 2004; Liu et al., 2013), using the so-called *kernel trick*.
4. **Multilayer Perceptrons (MLP):** Multilayer Perceptrons are powerful models that can model complex non-linear relationships between the input data and their targets using multiple layers (Haykin, 2009). Note that MLPs are often prone to overfitting, especially when a small number of training samples is used, leading to the development of several regularization methods, such as Dropout (Srivastava et al., 2014) (also abbreviated as “drop.” in this paper). The networks used in this paper are trained to minimize the squared error loss function given in (3) using the Adam optimizer (Kingma and Ba, 2014).

2.2. Class Decoding

The resulting regressed vectors are provided as input to a class decoder that selects the most probable image class. The original setup proposed by Horikawa and Kamitani (2017) employs a similarity metric to compare the decoded feature vector to each of the class representations. The class, whose representation is most similar to the regressed vector, is the best candidate image class as described in (2). In this paper we thoroughly study the effect of different similarity metrics on the accuracy of the decoding. The following similarity metrics are considered:

1. **Euclidean similarity:** The Euclidean similarity is computed as the inverse of Euclidean distance:

$$S_{euclidean}(\mathbf{y}_1, \mathbf{y}_2) = \frac{1}{1 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2}. \quad (4)$$

2. **Cosine similarity:** The cosine similarity is defined as the angle between two vectors:

$$S_{cosine}(\mathbf{y}_1, \mathbf{y}_2) = \frac{\mathbf{y}_1^T \mathbf{y}_2}{\|\mathbf{y}_1\|_2 \|\mathbf{y}_2\|_2}. \quad (5)$$

3. **Pearson similarity:** The Pearson similarity (correlation) between two vectors is computed as:

$$S_{pearson}(\mathbf{y}_1, \mathbf{y}_2) = \frac{(\mathbf{y}_1 - \boldsymbol{\mu}_1)^T (\mathbf{y}_2 - \boldsymbol{\mu}_2)}{\|\mathbf{y}_1 - \boldsymbol{\mu}_1\|_2 \|\mathbf{y}_2 - \boldsymbol{\mu}_2\|_2}, \quad (6)$$

where the notation $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is used to denote the average of the values in vectors \mathbf{y}_1 and \mathbf{y}_2 respectively.

2.3. Evaluation Protocol

Using the setup proposed in (Horikawa and Kamitani, 2017), the model’s accuracy is calculated by determining the percentage of the candidate categories for which the similarity between the given category and a test sample is smaller than the similarity with its correct class. However, this setup does not adequately reflect the decoding system’s accuracy, as it considers the prediction and the candidate classes in a pairwise (one-against-one) fashion. For every pair to be considered, the probability of a random match is 50%, as this setup only considers binary decision problems (correct class vs. one other candidate class). As a result, the identification accuracy reported using

this evaluation setup may be positively skewed by a significant amount, especially since in the original setup a large number of irrelevant distractors (with vastly different class representations) are used.

More often than not, rather than deciding between two candidate classes (one-against-one), we are interested in differentiating between multiple classes (one-against-all). To this end, we propose a more realistic evaluation setup where the category of each sample must be inferred among *all* the candidate categories. The most probable candidate image category is again inferred by measuring the similarity between the decoded feature vector and each of the class representation vector, as described in (2).

3. Experimental Evaluation

The experimental evaluation is provided in Section. First, the used dataset and feature extraction methods are described in Subsection 3.1, while the different experimental setups are introduced in Subsection 3.2. Finally, the experimental results are presented and discussed in Subsection 3.3.

3.1. Data

In the present study, we use the publicly available dataset described by Horikawa and Kamitani (2017). It consists of fMRI data from 5 human subjects performing visual tasks. Note that we cannot train a model using the data of all 5 subjects, since the fMRI representations extracted from different subjects can vary widely and performing transfer learning from one subject’s encoding to another’s is not trivial (Wen et al., 2018). Instead, a separate encoding model is trained and evaluated individually for every subject and the performance is averaged across all subjects in the final evaluation step. In total, for every subject, there exist 3,450 brain activity samples, 1,200 for training and 2,250 for testing, each corresponding to a visual stimulus. The stimuli images are natural images from the ImageNet dataset (Krizhevsky et al., 2012). Notably, the images that form the train and test sets belong to non-overlapping classes, i.e., ImageNet synsets, prohibiting the use of traditional classification algorithms for the task of neural decoding. In particular, the train set consists of 150 image classes and the test set of 50 image classes.

The data cover two distinct visual tasks performed by each subject: a) the image presentation task, where images were directly shown to the subjects and b) the imagery task, where the subjects were instructed to visualize images that correspond to a given word. The dataset contains a total of 1,750 test samples per subject for the image presentation experiments and 500 test samples per subject for the imagery experiments. The interested reader is referred to (Horikawa and Kamitani, 2017) for more details regarding the data collection protocol and experimental setups.

Each data sample contains information from multiple brain regions specialized in visual processing, that are summarized in Fig. 2. Specifically, the data include voxel information from:

- the areas V1, V2, V3 and V4, that are responsible for detecting edges, color and contours,

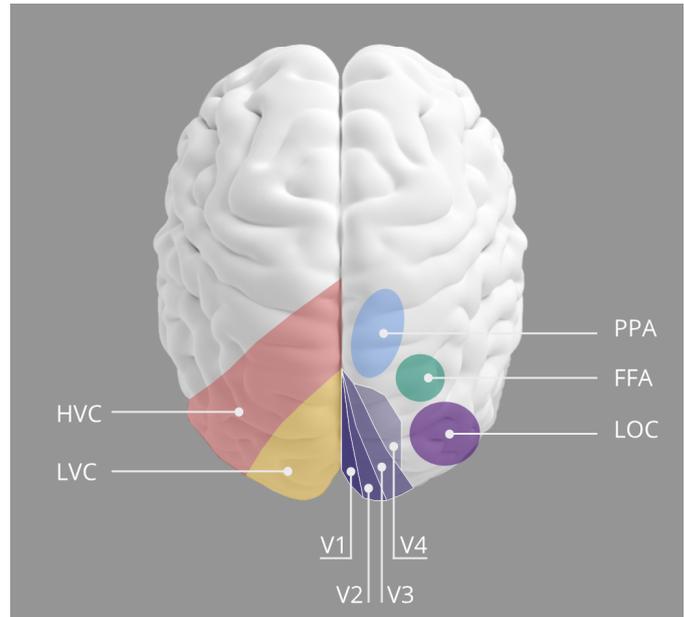


Fig. 2. Schematic view of the visual cortex with highlighted regions of interest

- the lateral occipital complex (LOC), that is involved with shape and object processing,
- the fusiform face area (FFA), that is specialized in the encoding and recognition of faces, and
- the parahippocampal place area (PPA), that is responsible for spatial information encoding.

Hence, three clusters of regions can be formed to provide a hierarchical understanding of the visual system: a) the lower visual cortex (LVC) that includes voxels from areas V1-V3, b) the higher visual cortex that comprises regions LOC, FFA, PPA and c) the entire visual cortex (VC) that encompasses all the aforementioned sub-regions. All the extracted voxels from the corresponding brain regions were used in the conducted experiments, i.e., we did not select only the voxels with the highest correlation, as in (Horikawa and Kamitani, 2017). This allows for retaining as much information as possible.

The images used as stimuli for the image presentation task are fed to a CNN that performs feature extraction. The use of a CNN layer for this task is not arbitrary. As discussed in (Horikawa and Kamitani, 2017), the features extracted from various layers of a CNN are tightly associated with the activity of various parts of the human visual cortex. The resulting feature vectors serve as the targets of the regression model that learns the mapping from recorded brain activity to natural images. Hence, every image is represented by an L -dimensional feature vector. In addition, for every class of images, a class representation is computed, as described in (1).

The AlexNet, which is composed of 5 convolutional (CNN1-CNN5) and 3 fully connected layers (FC1-FC3), was used to extract the target representations (Krizhevsky et al., 2012). The model was pre-trained on the ImageNet dataset (Krizhevsky et al., 2012), while 1,000 activations were randomly sam-

pled from each layer to compile the feature representation (Horikawa and Kamitani, 2017). For all reported experiments in this study, we used the available feature vectors, provided by the authors of (Horikawa and Kamitani, 2017) and produced as described above, to allow for easily reproducing the conducted experiments. Furthermore, we carried out multiple additional experiments using both image representations and class representations as targets.

3.2. Experimental Setups

The conducted experiments include extensive evaluation of:

- deep target features from different layers,
- multiple regression models,
- various similarity metrics,
- using image representations as the regressor’s targets compared to the class representations,
- the one-against-one versus the one-against-all approach for measuring the model’s accuracy.

Unless explicitly stated, all experiments use the feature representations extracted from layer CNN5 and are evaluated using the Pearson metric. Both the input features \mathbf{x}_i and the visual representations $\mathbf{y}_i^{(i)}$ were normalized using z-score normalization, i.e., normalized to have zero mean and unit variance. Note that this implies that the output of the regression model must be re-centered and re-scaled using the computed mean and standard deviations before performing the class decoding. The following hyper-parameters were used for all the conducted experiments. The 5 nearest neighbors were used for regressing the representations using the kNN model. The weight of the regularizer was set to 1 for the RR, while a 2nd degree polynomial kernel led to the best performance for the KR model (the weight of the regularizer was set to 0.005, while a constant value of 10 was added to the kernel). After experimenting with various MLP architectures, an MLP with with one hidden layer was used. The hidden layer is composed of 300 neurons that use sigmoid activation functions. The dropout probability was set to 30% (applied to the input layer only). The MLP models were trained for 100 epochs using batch size 128 employing the Adam optimizer (learning rate was set to 0.001 and the default hyper-parameters of the optimizer were used (Kingma and Ba, 2014)).

3.3. Experimental Results

The extensive evaluation of the decoding pipeline has yielded some interesting findings. The first set of experiments examines the effect of choosing different machine learning models for the regression tasks of image presentation and imagery. The same experiments were conducted using both the image-specific targets (Image Representation) and the class-averaged targets (Class Representation). A summary of the results is showcased in Fig. 3.

For the image presentation task, the best accuracy score is achieved by the MLP regressor for the VC brain region. The predominance of the MLP model becomes even more evident

when using the class-averaged representations (Fig. 3.c). Regions V1-V4 seem to perform increasingly better, a finding that reflects their inherent hierarchical structure in the brain itself. Another interesting finding is that for the FFA region, non-linear models outperform the linear models, while for the PPA region the opposite is true. HVC’s performance exceeds the performance of each of its sub-regions. This finding suggests, that the voxels of the higher visual cortex areas provide complementary information, leading to much richer representations when processed as one structure. However, there is a drop in the accuracy score for all brain regions when the models were trained with the class representations. This may be explained by the fact that image representations preserve the features’ diversity power, as opposed to the class representations that are averaged, and hence collapsed in the feature space.

Regarding the imagery task, this trend is reversed. Using class representations yield significantly better results than their image-specific counterparts (Fig. 3.d). Indeed, the imagery task experimental protocol required that the subject visualized on cue as many images of a specific category as possible. Thus, the internal thought process of the subject might approximate the averaged class representation, explaining the better performance of these targets. This time, the top performing models are kNN and MLP+drop., both in the HVC region, for the image-specific and class-averaged targets respectively. In general, kNN seems to be more robust to distribution shifts caused by feeding the brain activity features of the imagery task to the model, since its output is always a constrained linear combination of the training targets.

Overall, the accuracy achieved in the imagery task is critically lower than the image presentation task. We speculate that this is the case because the imagery task carries a greater intrinsic variability, as the high-level encoding circuitry is not as genetically hard-coded as the primary visual cortex and its adjacent sub-regions. The latter structures, are present in all mammals (Krubitzer, 2007), have been preserved for millions of years across species and, therefore, their function is far more homogeneous among humans compared to higher visual areas.

Our experimental results on the correspondence of certain brain regions with the levels of deep CNN representations, support the finding of (Horikawa and Kamitani, 2017) that suggests tight association between hierarchical visual areas and the complexity levels of visual features. Indeed, as illustrated in Fig. 4.a, for the image presentation task, features extracted from the convolutional layers tend to be better predicted by brain regions of the lower visual cortex. However, for the imagery task (Fig. 4.b), higher cortex brain regions outperform the lower visual cortex sub-regions for all level representations, as was also the case in Fig. 3.b and 3.d. If we assume that *the internal processes of the two tasks in the brain are similar*, this finding suggests that the lower visual cortex areas are not particularly involved during visual imagery tasks. This hypothesis is also supported by the fact that HVC alone is the top performing region for the imagery task, compared to the presentation task where VC, encompassing all sub-regions, scored the highest. If this hypothesis is false, then these results may hint that the analogy between CNNs and cortical hierarchy does not hold

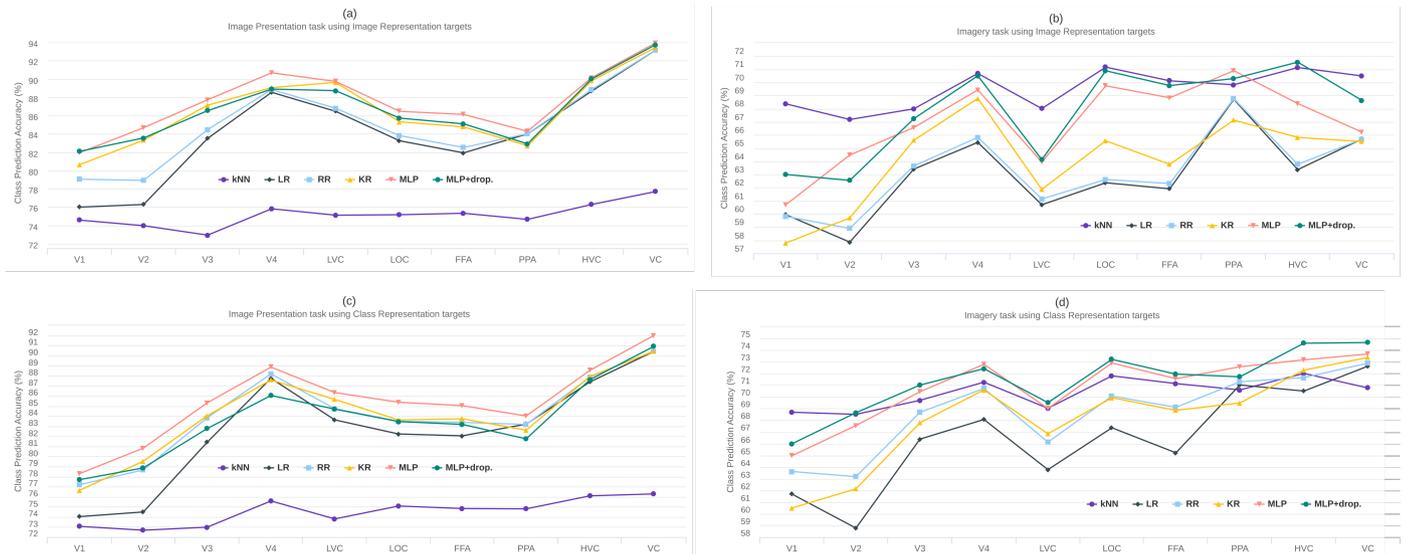


Fig. 3. Impact of various regression models on decoding accuracy

for tasks other than the image presentation. Nevertheless, the results clearly point to certain correlations between deep representations and brain regions, a property that can be further exploited to obtain better accuracy.

The different similarity metrics (described in Section 2.3) were evaluated for the best performing model of both the image presentation and imagery tasks, i.e., for the MLP and kNN models respectively. As shown in Fig. 5, Pearson similarity yields the best results for the image presentation task, while Euclidean similarity considerably boosts the accuracy score on the imagery task.

Finally, the evaluation results for the one-against-one and one-against-all approaches is depicted in Fig. 6 for the same model, brain region and deep representation. The results refer to the problem of generic decoding from 50 unseen image categories, hence the one-against-all chance level is 2%. Note that the accuracy scores drop significantly when the one-against-all method is applied, highlighting the importance of following a robust protocol for the evaluation process. Nonetheless, the results for both tasks are analogous to the previous findings regarding the margin from chance level. These results suggest that we are still far away from developing practical and robust methods for performing visual representation decoding from human brain activity, while there is much room for improvement of the existing methods for the aforementioned task.

For a complete report of the experimental results, that can be used for comparing new methods to the presented baselines and generating accuracy plots, the reader is referred to the Supplementary Material.

4. Conclusions

An extensive evaluation of the decoding pipeline was performed in this paper, experimenting with various parts of the process, from the appropriate selection of the target features and regression models to the similarity metrics used in the class

decoding step. Furthermore, a more general one-against-all approach was introduced for the evaluation step, to more reliably capture the system's accuracy.

The reported experimental evaluation is expected to serve as a baseline for the task of visual representation decoding. This work provides empirical information about which models perform best for each visual task, how to determine which feature layer representation to use as the regressor's targets and what similarity metric most efficiently captures the complex relationships of the decoded image representations. Interestingly, the reported results may also provide some insight regarding our own brain function on visual processing, as is hinted in Section 3.

References

- Amin, J., Sharif, M., Yasmin, M., Fernandes, S.L., 2017. A distinctive approach in brain tumor detection and classification using mri. *Pattern Recognition Letters*.
- Arunkumar, N., Ramkumar, K., Venkatraman, V., Abdulhay, E., Fernandes, S.L., Kadry, S., Segal, S., 2017. Classification of focal and non focal eeg using entropies. *Pattern Recognition Letters* 94, 112–117.
- Cruz-Garza, J.G., Hernandez, Z.R., Nepal, S., Bradley, K.K., Contreras-Vidal, J.L., 2014. Neural decoding of expressive human movement from scalp electroencephalography (eeg). *Frontiers in Human Neuroscience* 8, 188.
- Cui, J., Zhu, Q., Wang, D., Li, Z., 2018. Learning robust latent representation for discriminative regression. *Pattern Recognition Letters*.
- Domingues, M.A., de Souza, R.M., Cysneiros, F.J.A., 2010. A robust method for linear regression of symbolic interval data. *Pattern Recognition Letters* 31, 1991–1996.
- Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., van Gerven, M.A., 2017. Reconstructing perceived faces from brain activations with deep adversarial neural decoding, in: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 4246–4257.
- Gwin, J.T., Gramann, K., Makeig, S., Ferris, D.P., 2010. Removal of movement artifact from high-density eeg recorded during walking and running. *Journal of Neurophysiology* 103, 3526–3534.
- Hastie, T., Tibshirani, R., 1996. Discriminant adaptive nearest neighbor classification and regression, in: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 409–415.
- Haykin, S.S., 2009. *Neural networks and learning machines*. volume 3. Pearson Education Upper Saddle River.

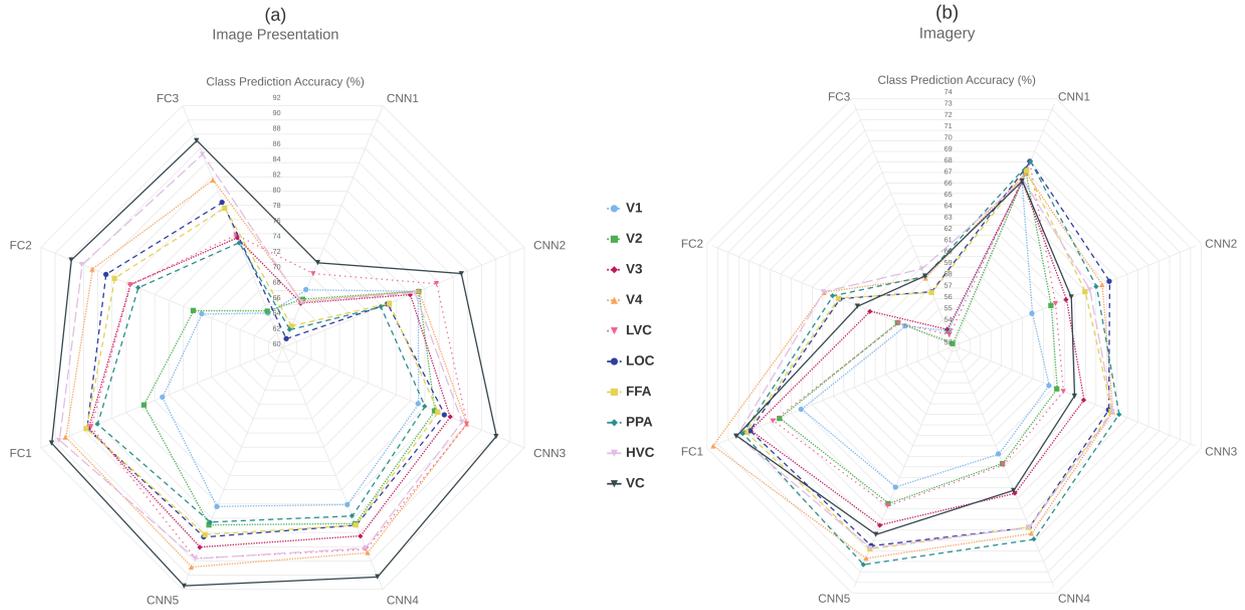


Fig. 4. Effect of using deep features from different layers for the (a) Image Presentation and (b) Imagery tasks

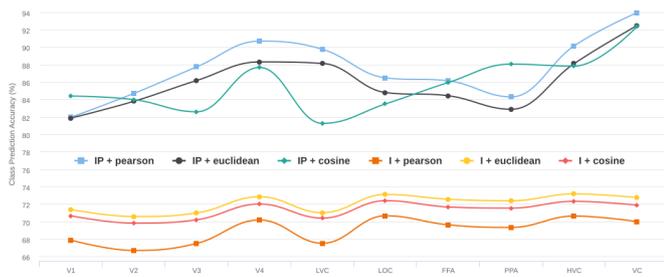


Fig. 5. Effect of different similarity metrics on the evaluation step for the Image Presentation (IP) and Imagery (I) tasks

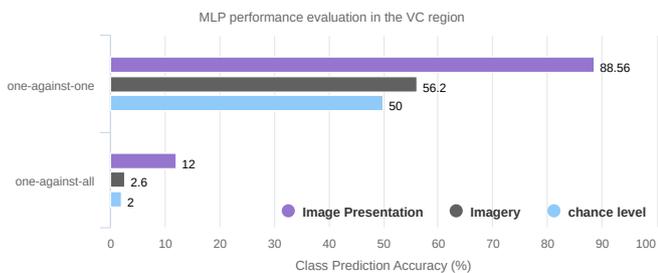


Fig. 6. Comparison of one-against-one and one-against-all approaches

- Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.
- Horikawa, T., Kamitani, Y., 2017. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications* 8, 15037.
- Horikawa, T., Tamaki, M., Miyawaki, Y., Kamitani, Y., 2013. Neural decoding of visual imagery during sleep. *Science* 340, 639–642.
- Huettel, S.A., Song, A.W., McCarthy, G., et al., 2004. *Functional magnetic resonance imaging. volume 1.* Sinauer Associates Sunderland, MA.
- Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization, in: *Proceedings of the International Conference on Learning Representations.*
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1097–1105.
- Krubitzer, L., 2007. The magnificent compromise: cortical field evolution in mammals. *Neuron* 56, 201–208.
- Liu, J., Zhao, F., Liu, Y., 2013. Learning kernel parameters for kernel fisher discriminant analysis. *Pattern Recognition Letters* 34, 1026–1031.
- Mosteller, F., Tukey, J.W., 1977. *Data analysis and regression: a second course in statistics.* Addison-Wesley Series in Behavioral Science: Quantitative Methods .
- Nasrabi, N.M., 2007. Pattern recognition and machine learning. *Journal of Electronic Imaging* 16, 049901.
- Niedermeyer, E., da Silva, F.L., 2005. *Electroencephalography: basic principles, clinical applications, and related fields.* Lippincott Williams & Wilkins.
- Pacola, E., Quandt, V., Liberalesso, P., Pichorim, S., Schneider, F., Gamba, H., 2017. A versatile eeg spike detector with multivariate matrix of features based on the linear discriminant analysis, combined wavelets, and descriptors. *Pattern Recognition Letters* 86, 31–37.
- Patel, R., Gireesan, K., Sengottuvel, S., 2018. Decoding non-linearity for effective extraction of the eye-blink artifact pattern from eeg recordings. *Pattern Recognition Letters* .
- Plumpton, C.O., 2014. Semi-supervised ensemble update strategies for on-line classification of fmri data. *Pattern Recognition Letters* 37, 172–177.
- Presacco, A., Goodman, R., Forrester, L., Contreras-Vidal, J.L., 2011. Neural decoding of treadmill walking from noninvasive electroencephalographic signals. *Journal of Neurophysiology* 106, 1875–1887.
- Schütze, H., Manning, C.D., Raghavan, P., 2008. *Introduction to information retrieval. volume 39.* Cambridge University Press.
- Shawe-Taylor, J., Cristianini, N., et al., 2004. *Kernel methods for pattern analysis.* Cambridge university press.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1929–1958.
- Tagliazucchi, E., Laufs, H., 2014. Decoding wakefulness levels from typical fmri resting-state data reveals reliable drifts between wakefulness and sleep. *Neuron* 82, 695–708.
- Wen, H., Shi, J., Chen, W., Liu, Z., 2018. Transferring and generalizing deep-learning-based neural encoding models across subjects. *NeuroImage* 176, 152–163.
- Wen, H., Shi, J., Zhang, Y., Lu, K.H., Cao, J., Liu, Z., 2017. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex* , 1–25.
- Yuan, J., Wang, D., Li, R., 2012. Image segmentation using local spectral histograms and linear regression. *Pattern Recognition Letters* 33, 615–622.